A Collection of Benchmark data sets for Knowledge Graph-based Similarity in the Biomedical Domain

Carlota Cardoso[1], Rita T. Sousa[1], Sebastian Köhler[2], Catia Pesquita[1]

[1] LASIGE, Faculdade de Ciências da Universidade de Lisboa, Portugal
{cmacardoso ✉ ,risousa,clpesquita}@ciencias.ulisboa.pt
[2] Monarch Initiative, monarchinitiative.org

Abstract. The ability to compare entities within a knowledge graph is a cornerstone technique for several applications, ranging from the integration of heterogeneous data to machine learning. It is of particular importance in biomedical applications such as prediction of protein-protein interactions, associations between diseases and genes, cellular localization of proteins, among others. However, building a gold standard data set to support their evaluation is non-trivial, due to size, diversity and complexity of biomedical knowledge graphs.
We present a collection of 21 benchmark data sets that aim at circumventing the difficulties in building benchmarks for large biomedical knowledge graphs by exploiting proxies for biomedical entity similarity. These data sets include data from two successful biomedical ontologies, the Gene Ontology and the Human Phenotype Ontology, and explore proxy similarities based on protein and gene properties. Data sets have varying sizes and cover four different species at different levels of annotation completion. For each data set we also provide semantic similarity computations with state of the art representative measures. Available at: https://github.com/liseda-lab/kgsim-benchmark

1 Introduction

Nearly all domains of human endeavour are responsible for producing large amounts of complex data, and the life sciences domain is a good example of this: high throughput techniques in genomics and proteomics produce large amounts of complex and unstructured data about the function, regulation and interaction of genes and proteins. This was a strong motivator for the adoption of ontologies, as the result of tremendous effort to make data understandable by both humans and machines. The ability to describe complex entities resorting to ontologies supports the computation of semantic similarity (SS) between entities. Several tasks can be supported by these metrics, in fact, entity similarity is an integral part of many machine learning techniques. For instance, SS has been successfully applied to prediction of protein-protein interaction [8] and clustering [5]. A semantic similarity measure (SSM) is a function that, given two ontology classes or two sets of classes describing two individuals, returns a numerical value reflecting the closeness in meaning between them [11]. There are several measures available [11] and each formalizes similarity in a slightly different way. However, determining the best measure for each application is still an open question since there is no gold standard. One possible solution is to compare SSMs to proxies of similarity that compare entities through different lenses. For instance, two proteins can be compared via their sequence, structure or common metabolic pathways. These can be used to assess how well SSMs capture entity similarity.

We present a collection of 21 benchmark data sets that aim at circumventing the difficulties in building benchmarks for large biomedical knowledge graphs (KGs) by exploiting proxies for biomedical entity similarity. These data sets are grouped according to the KG and proxy measures they are based on and have a wide range of sizes, from a few hundreds to over 150 thousand pairs.

2 Related Work

Building a gold standard data set to support semantic similarity evaluation is not trivial. Accomplishing this manually is extremely time consuming and existing manual gold standards are very small compared to the size of the ontologies they correspond to. For instance, Pedersen et al. [9] created a set of only 30 term pairs extracted from a universe of over 1 million of biomedical concepts from UMLS. To mitigate this challenge, some semantic web related applications have turned to crowd-sourcing (e.g. ontology matching [4]), which brings with it a series of new challenges. The evaluation task can be inherently biased towards a particular viewpoint of the domain and is highly dependent on the ability to provide crowd-sourced workers with enough information to make a decision.

In previous works, we developed CESSM [12], a tool for the evaluation of new SSMs through comparison with previously published ones, and considering their relation to different similarity proxies. CESSM has been adopted by the community and used to evaluate several novel SSMs. Overtime some limitations of its use have been identified, namely being limited to a single ontology and being focused on a single functional perspective (molecular function similarity).

Finally, there are related contributions in the area of benchmark data for link prediction [3] and classification in KGs [13]. KG-based SS can be applied in these contexts, but these benchmark data sets do not support a direct evaluation of SSMs.

3 Building the benchmark data sets

Each benchmark data set is made up of several pairs of biomedical entities (e.g. proteins or genes) and their respective state of the art SS and proxy similarity values. In building them, the first step is to select features for the entities and pairs. Entities should be as thoroughly described within the context of the ontology as possible, while pairs should capture the full spectrum of similarity values throughout.

Calculating the SS between entities described by sets of ontology classes usually combines a measure of the information content (IC) of a class (i.e. a measure of its specificity) and an approach to calculate similarity between all the classes. 2 different approaches for the calculation of entity SS were combined with 2 methods for IC calculation [11] to arrive at four state of the art SSMs employed in the data sets: BMA_{Resnik}, BMA_{Seco}, $simGIC_{Resnik}$ and $simGIC_{Seco}$.

3.1 Protein benchmark data sets

The majority of research into SSMs in bioinformatics is applied to the Gene Ontology (GO) [10], the most widely adopted ontology by the life sciences community, which covers three distinct aspects of gene product's roles: molecular function, cellular component and biological process [2]. We built two types of protein benchmark data sets: one based on molecular functional similarity and another based on protein-protein interactions (PPI). To ensure enough information, the data sets are constituted by protein pairs, in which each protein is sufficiently annotated with GO classes and with the necessary information to compute proxy similarity[3]. This results in two annotation levels: *One aspect*, where all proteins are well annotated in at least one GO aspect; and its subset, *All aspects*, where all proteins are well annotated in all GO aspects.

We employ three proxies of protein similarity based on their biological properties: (1) **sequence similarity** measures the relationship between two protein sequences; (2) **molecular function similarity** compares the functional regions that exist in each protein sequence; and (3) **protein-protein interaction** that determines if the proteins interact or not. In the molecular function similarity data sets, we employ (1) and (2), whereas in the protein-protein interactions data sets, we employ (1) and (3). The proposed methodology was employed to produce the data sets described in Table 1.

Table 1. Summary of the Protein benchmarks

	Protein-Protein Interaction				Molecular Function			
	One Aspect		*All Aspects*		*One Aspect*		*All Aspects*	
Species	Proteins	Pairs	Proteins	Pairs	Proteins	Pairs	Proteins	Pairs
D. melanogaster	481	397	335	270	7.494	53.797	5.810	52.460
E. coli	371	738	264	428	1.250	4.623	748	1.813
H. sapiens	7.644	44.677	7.149	42.204	1.3604	60.176	12.487	60.163
S. cerevisae	3.874	34.772	2.959	21.577	4.783	42.192	3.660	30.747
All	12.370	80.584	10.707	64.479	27.131	158.512	22.705	142.736

3.2 Genes-phenotypes benchmark data set

The Human Phenotype Ontology (HPO), another widely used biomedical ontology, provides a standardized vocabulary of phenotypic abnormalities encoun-

[3] PPI information retrieved from the data sets used in [14]. Molecular function similarity is based on Pfam [6] assignments to proteins.

Table 2. Pearson correlation coefficient between semantic similarity ($simGIC_{Seco}$) and biological similarity proxies for *H.sapiens*

	One Aspect		All Aspects	
Data Set	sim_{Seq}	sim_{Pfam}	sim_{Seq}	sim_{Pfam}
Molecular Function	0.723	0.612	0.732	0.620
PPI	0.536	0.422	0.546	0.421
Disease-Phenotype	0.482			

tered in human diseases [7]. The HPO KG integrates the links between human genes and their associated HPO classes. Human genes without sufficient and specific annotations, or the necessary information to compute proxy similarity were filtered. The similarity proxy selected for this data set is **phenotypic series similarity**, computed by comparing the phenotypic series (groups of identical or similar phenotypes [1]) related to each gene. After selecting the eligible genes, pairs were filtered to ensure the same number of pairs with null, not-null and full phenotypic series similarity. Following this methodology resulted in a data set with 2026 distinct human genes and 12000 pairs.

Table 2 shows correlation values for all *H. sapiens* data sets with $simGIC_{Seco}$.

4 Conclusions

The collection of benchmark data sets we present aims at supporting the large-scale evaluation of KG-based SS. All data sets and KG data used to compute the SSMs are available online[4]. This allows for a direct comparison with the pre-computed semantic similarity measures, as well as facilitates the direct comparison between different works using this resource. For this reason, the benchmark will purposefully remain static for a few years, following the approach used by CESSM [12], released in 2009 and updated in 2014. Parallel updates to the benchmark data sets will include new KG, with updated attributes for entity selection and new similarity proxies.

The benchmark supports simple evaluation metrics, such as computing Pearson's correlation coefficient between the SSMs and the similarity proxies, but it also supports more complex evaluations. For instance, the PPI data sets also support prediction of protein-protein interaction based on semantic similarity, as done in Sousa et al. [14]. Despite being domain-specific, we expect this collection to be useful beyond the biomedical domain. Similarity computation within KG is a fundamental building block of many semantic web applications ranging from data integration to data mining, meaning the benchmark data sets can be used for the evaluation of SSMs developed outside the biomedical domain. Alternatively, the general approach developed for the creation of the data sets is generalizable to any domain where a similarity proxy can be created, making the development of analogous benchmark data sets outside the biomedical domain

[4] `https://github.com/liseda-lab/kgsim-benchmark`

a possibility.

Acknowledgements: CP, CC and RTS were funded by the Portuguese FCT through the LASIGE Research Unit (UIDB/00408/2020 and UIDP/00408/2020), and also by the SMILAX project (PTDC/EEI-ESS/4633/2014). RTS was also funded by SFRH/BD/145377/2019.

References

1. Amberger, J.S., Bocchini, C.A., Schiettecatte, F., Scott, A.F., Hamosh, A.: Omim. org: Online mendelian inheritance in man (omim®), an online catalog of human genes and genetic disorders. Nucleic acids research **43**(D1), D789–D798 (2014)
2. Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., et al.: Gene ontology: tool for the unification of biology. Nature Genetics **25**(1), 25–29 (2000)
3. Bordes, A., Usunier, N., Garcia-Duran, A., Weston, J., Yakhnenko, O.: Translating embeddings for modeling multi-relational data. In: Advances in neural information processing systems. pp. 2787–2795 (2013)
4. Cheatham, M., Hitzler, P.: Conference v2. 0: An uncertain version of the oaei conference benchmark. In: International Semantic Web Conference. pp. 33–48 (2014)
5. Chen, J., Liu, Y., Sam, L., Li, J., Lussier, Y.: Evaluation of high-throughput functional categorization of human disease genes. BMC bioinformatics **8 Suppl 3**, S7 (02 2007)
6. El-Gebali, S., Mistry, J., Bateman, A., Eddy, S.R., Luciani, A., et al.: The Pfam protein families database in 2019. Nucleic Acids Research **47**(D1), D427–D432 (10 2018)
7. Köhler, S., Carmody, L., Vasilevsky, N., Jacobsen, J.O.B., Danis, D., et al.: Expansion of the Human Phenotype Ontology (HPO) knowledge base and resources. Nucleic Acids Research **47**(D1), D1018–D1027 (11 2018)
8. Maetschke, S.R., Simonsen, M., Davis, M.J., Ragan, M.A.: Gene ontology-driven inference of protein–protein interactions using inducers. Bioinformatics **28**(1), 69–75 (2011)
9. Pedersen, T., Pakhomov, S.V., Patwardhan, S., Chute, C.G.: Measures of semantic similarity and relatedness in the biomedical domain. Journal of biomedical informatics **40**(3), 288–299 (2007)
10. Pesquita, C.: Semantic similarity in the gene ontology. In: Dessimoz, C., Škunca, N. (eds.) The Gene Ontology Handbook, pp. 161–173. Humana Press, New York, NY, USA (2017)
11. Pesquita, C., Faria, D., Falcao, A.O., Lord, P., Couto, F.M.: Semantic similarity in biomedical ontologies. PLoS computational biology **5**(7) (2009)
12. Pesquita, C., Pessoa, D., Faria, D., Couto, F.: Cessm: collaborative evaluation of semantic similarity measures. JB2009: challenges in bioinformatics **157**, 190 (2009)
13. Ristoski, P., de Vries, G.K.D., Paulheim, H.: A collection of benchmark datasets for systematic evaluations of machine learning on the semantic web. In: The Semantic Web – ISWC 2016. pp. 186–194. Springer International Publishing, Cham, Switzerland (2016)
14. Sousa, R.T., Silva, S., Pesquita, C.: Evolving knowledge graph similarity for supervised learning in complex biomedical domains. BMC Bioinformatics **21**(1) (2020)